Perspective Homework #3

Yuqing Zhang

5/14/2017

Problem 1

```
biden_data<-read_csv('biden.csv')  %>%
  mutate(obs_num = as.numeric(rownames(.))) %>%
  mutate(dem = factor(dem),
         rep = factor(rep))
```

```
## Parsed with column specification:
## cols(
##   biden = col_integer(),
##   female = col_integer(),
##   age = col_integer(),
##   educ = col_integer(),
##   dem = col_integer(),
##   rep = col_integer()
## )
```

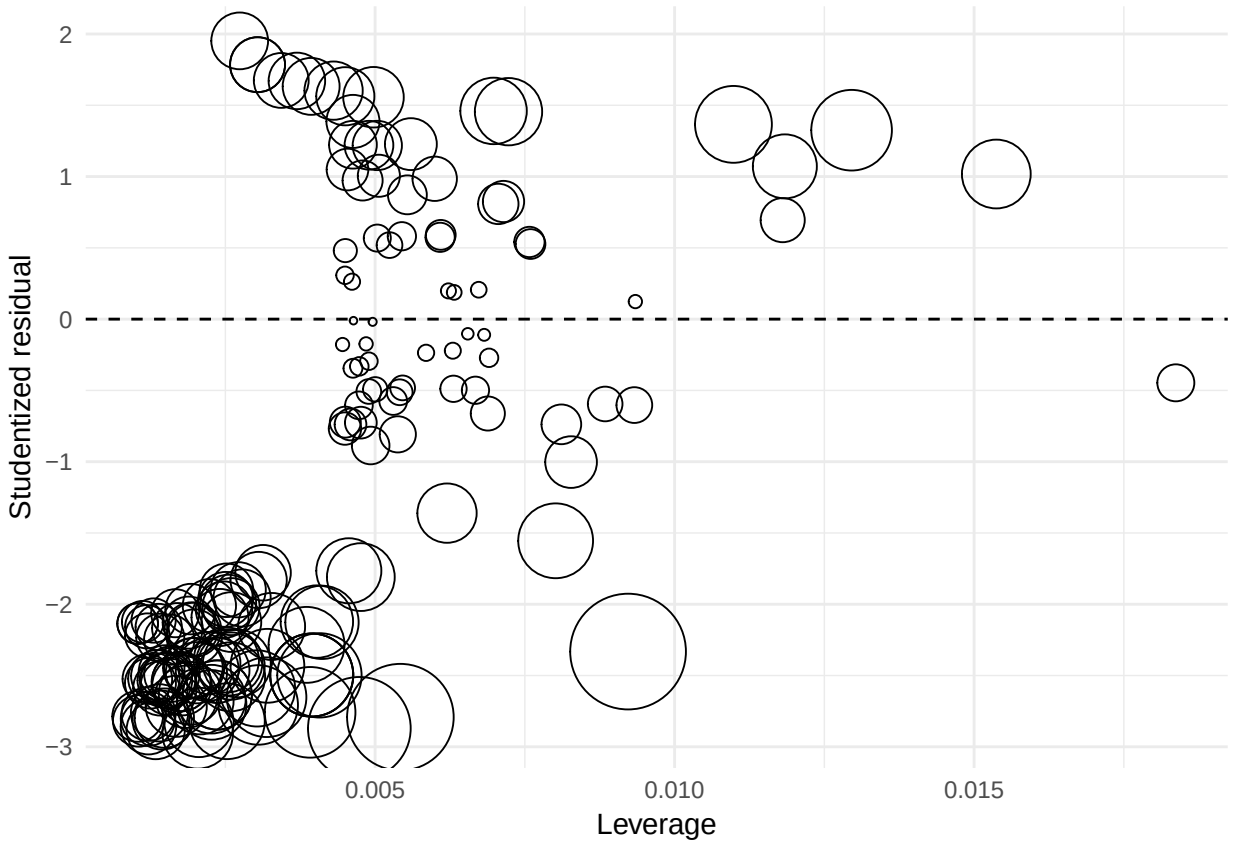
```
biden <- biden_data %>%
  na.omit()
biden_mod <- lm(biden ~ age+educ+female,data=biden)
tidy(biden_mod)
```

```
##          term estimate std.error statistic  p.value
## 1 (Intercept)  68.6210    3.5960     19.08 4.34e-74
## 2         age   0.0419    0.0325      1.29 1.98e-01
## 3        educ  -0.8887    0.2247     -3.96 7.94e-05
## 4      female   6.1961    1.0967      5.65 1.86e-08
```

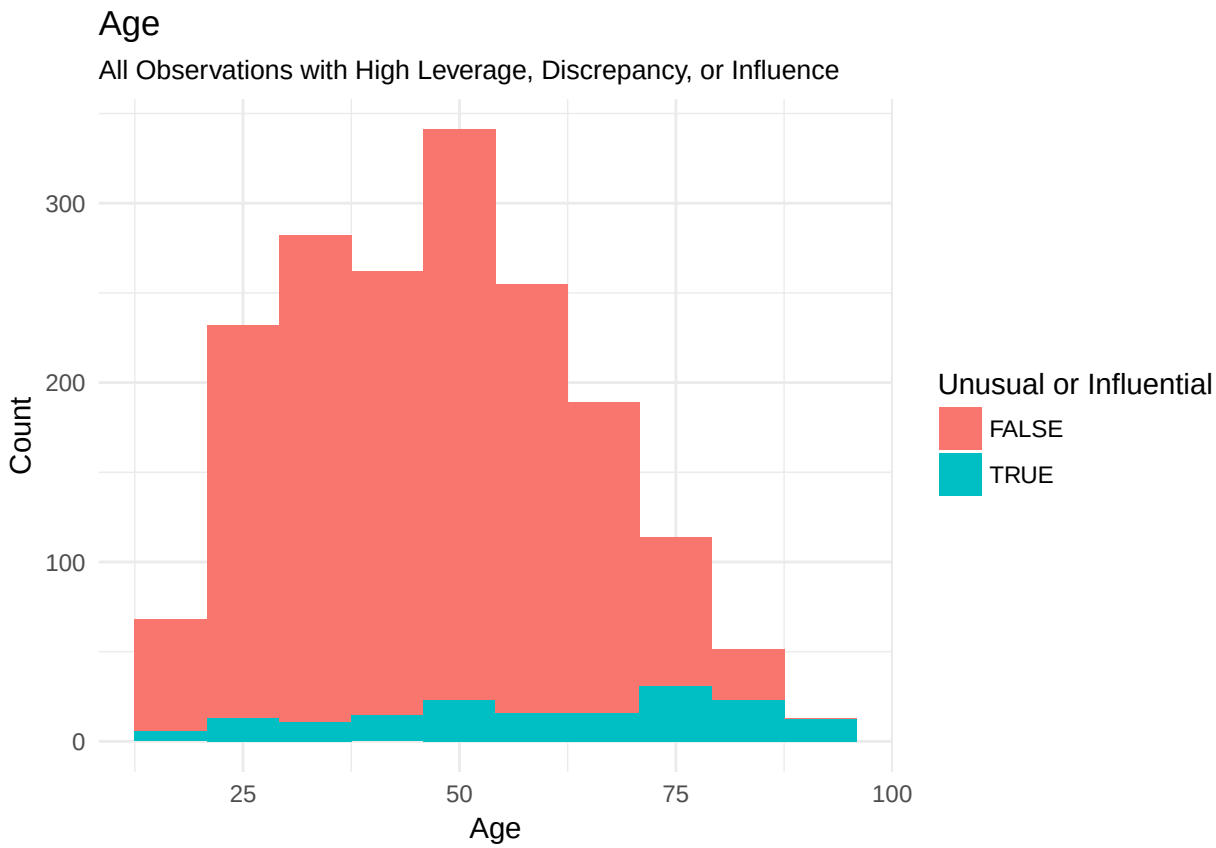
The coefficients for age, education and female are: 0.0419,-0.8887,6.1961 and the standard errors for age, education and female are:0.0325,0.2247,1.0967.

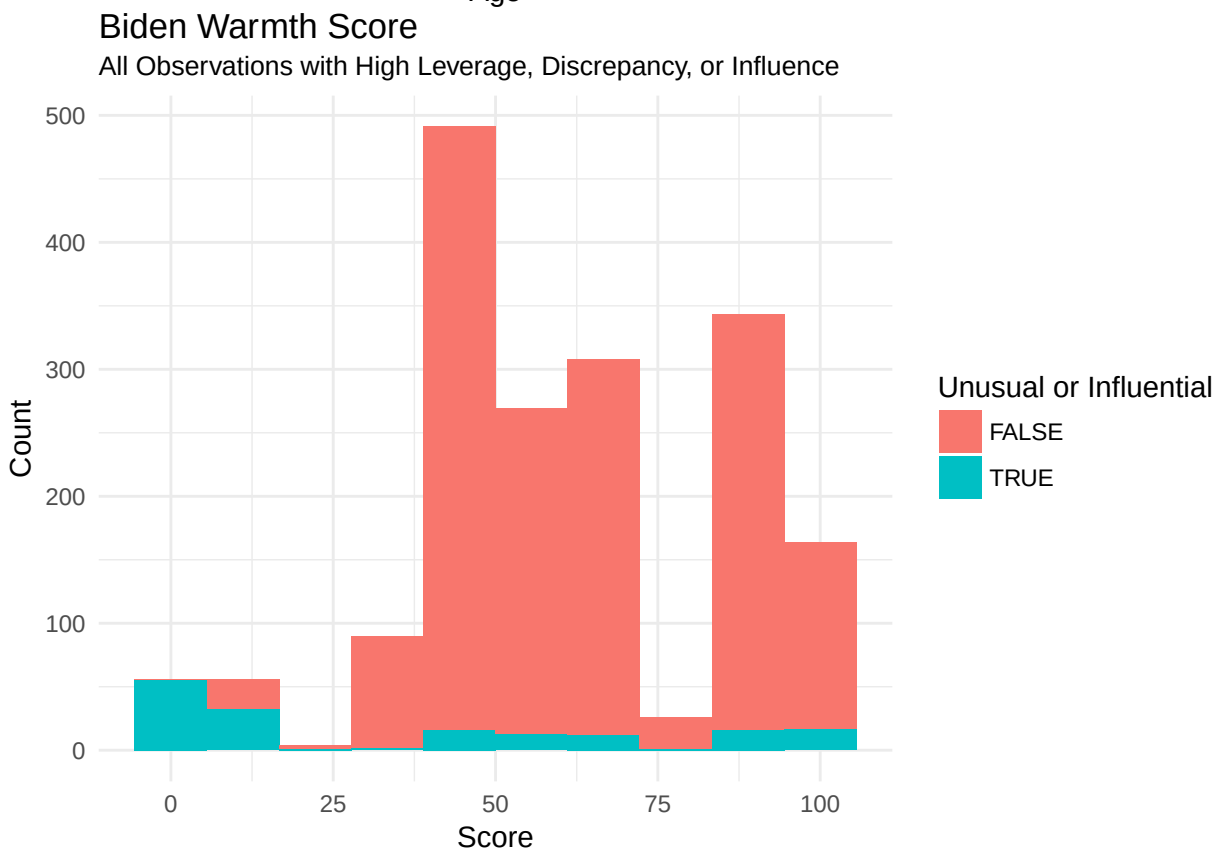
1. Test the model to identify any unusual and/or influential observations.

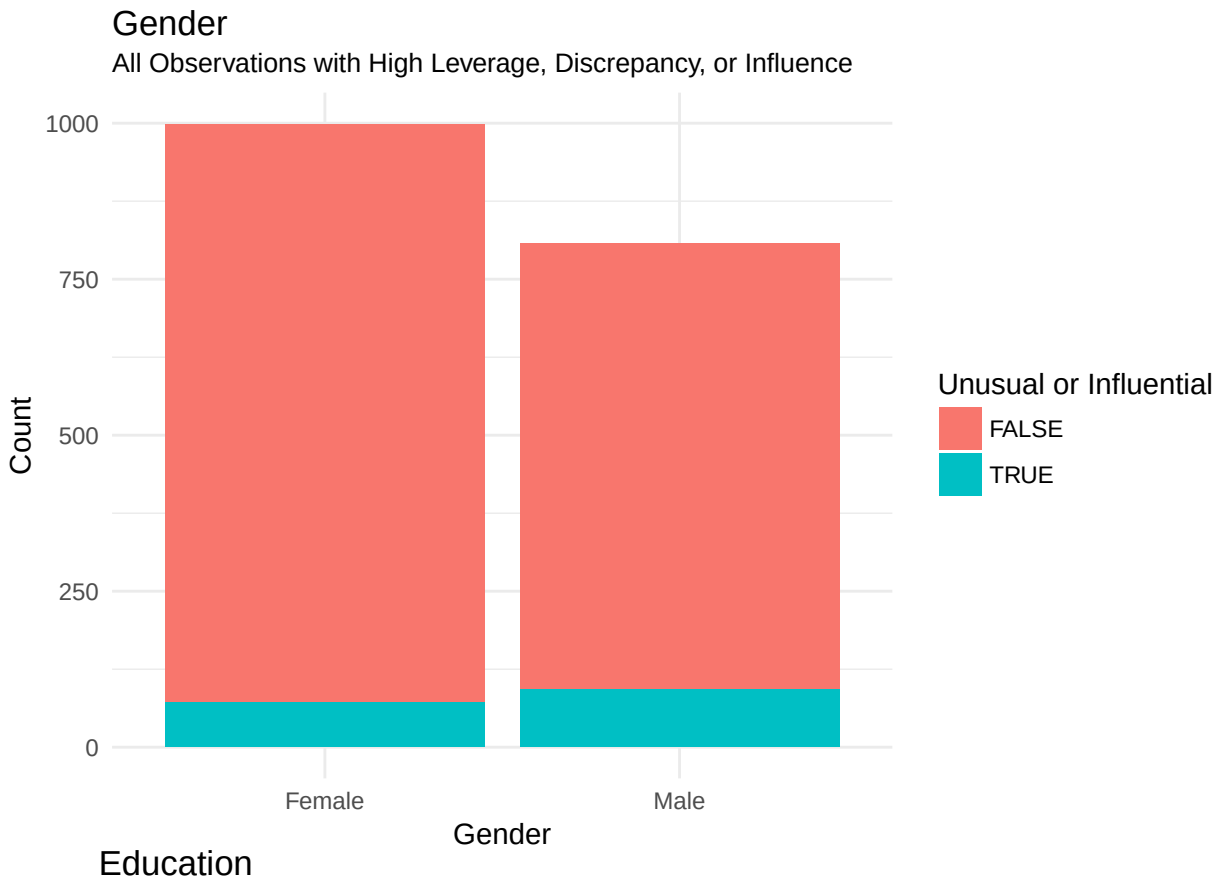
Let's use a bubble plot to identify any unusual or influential observations.

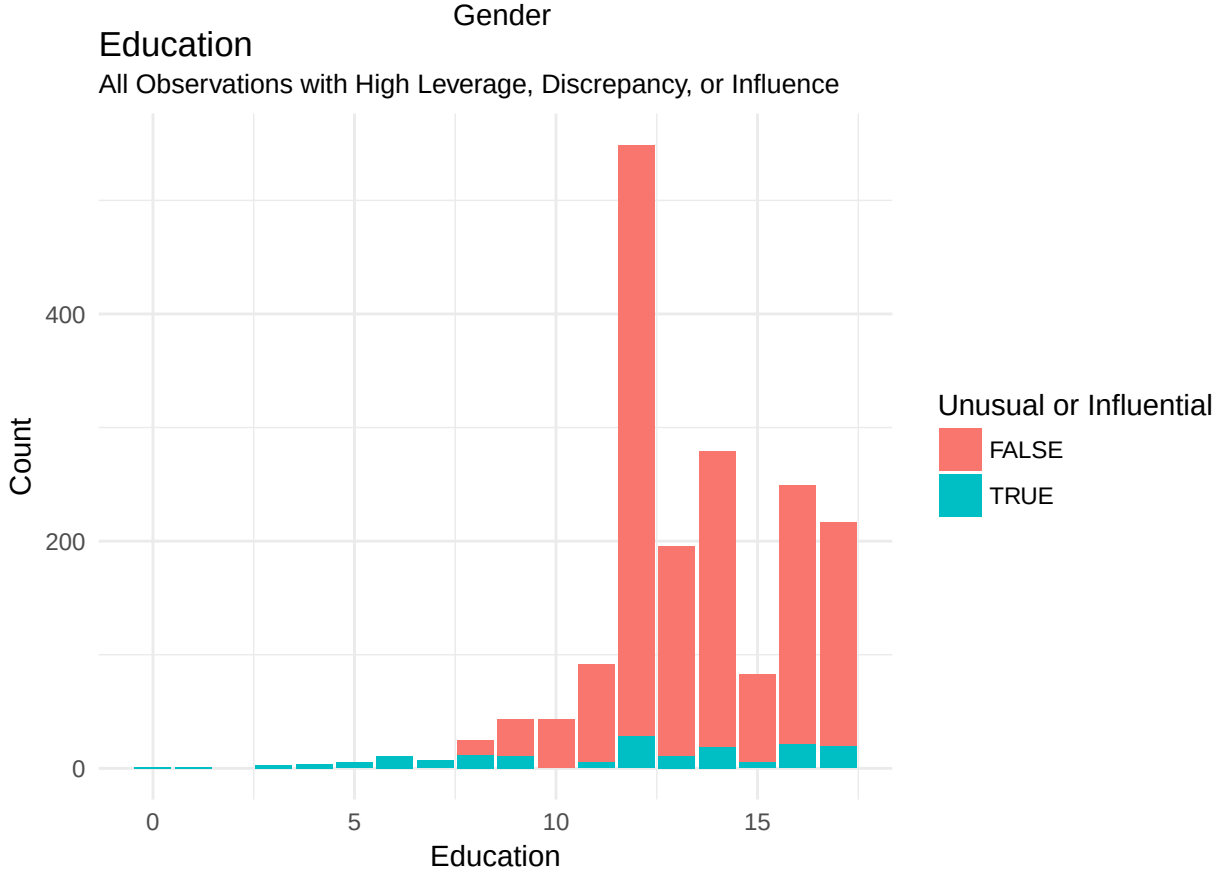


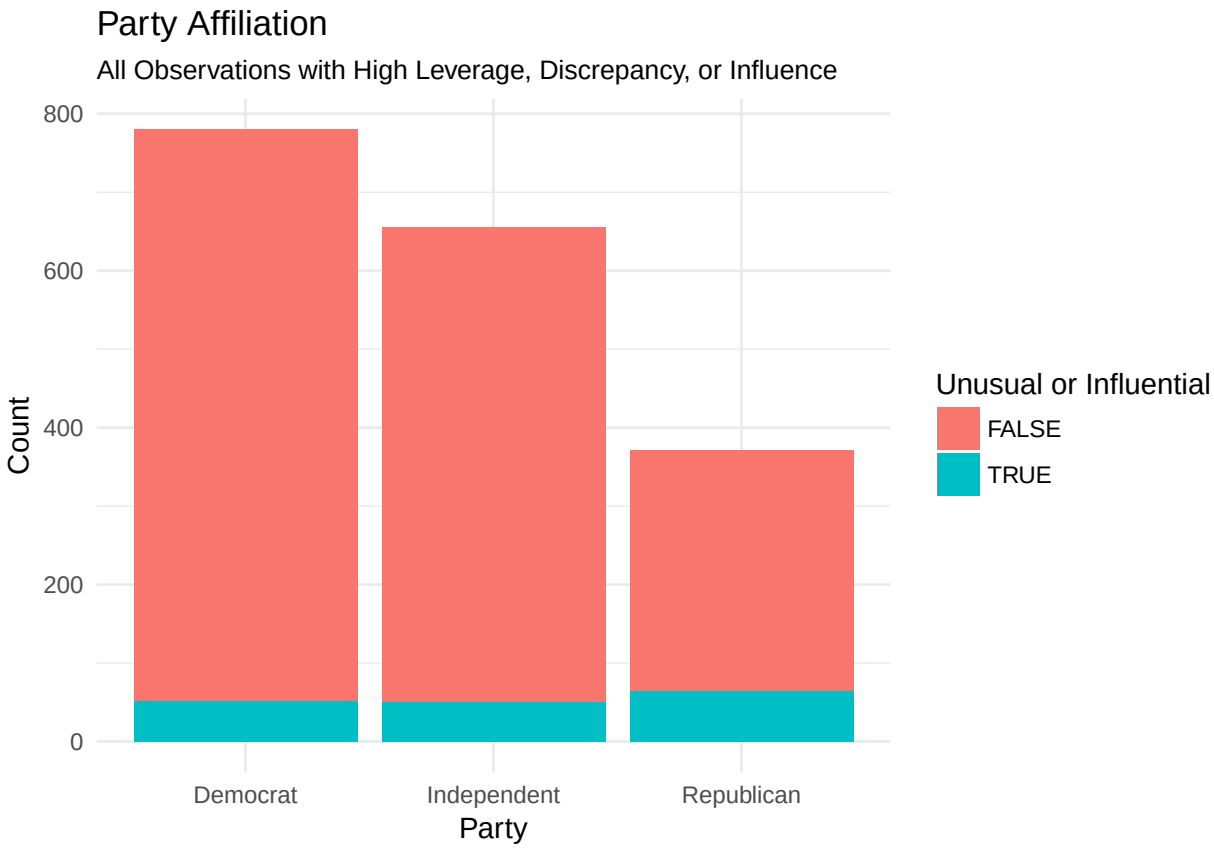
From the above bubble plot we can see there are 167 observations that are unusual and influential. Most of them are located at the lower left part of the plot,meaning that they have high discrepancy but lower leverage level. So what variables are representative in causing unsual or influential?


Age
All Observations with High Leverage, Discrepancy, or Influence
Count
300
200
100
0
25
50
75
100
Age
Unusual or Influential
FALSE
TRUE


Biden Warmth Score
All Observations with High Leverage, Discrepancy, or Influence
Count
500
400
300
200
100
0
0
25
50
75
100
Score
Unusual or Influential
FALSE
TRUE

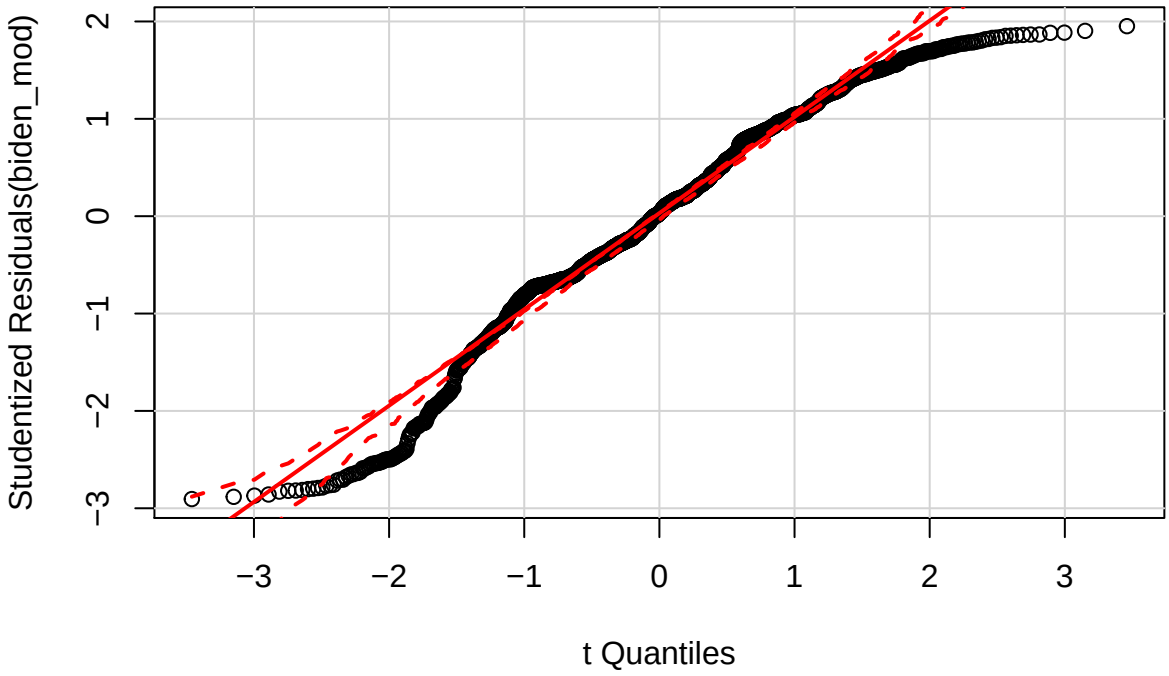

Gender
All Observations with High Leverage, Discrepancy, or Influence
1000
750
500
250
0
Count
Female
Male
Gender
Unusual or Influential
FALSE
TRUE


Education
All Observations with High Leverage, Discrepancy, or Influence
400
200
0
Count
0
5
10
15
Education
Unusual or Influential
FALSE
TRUE

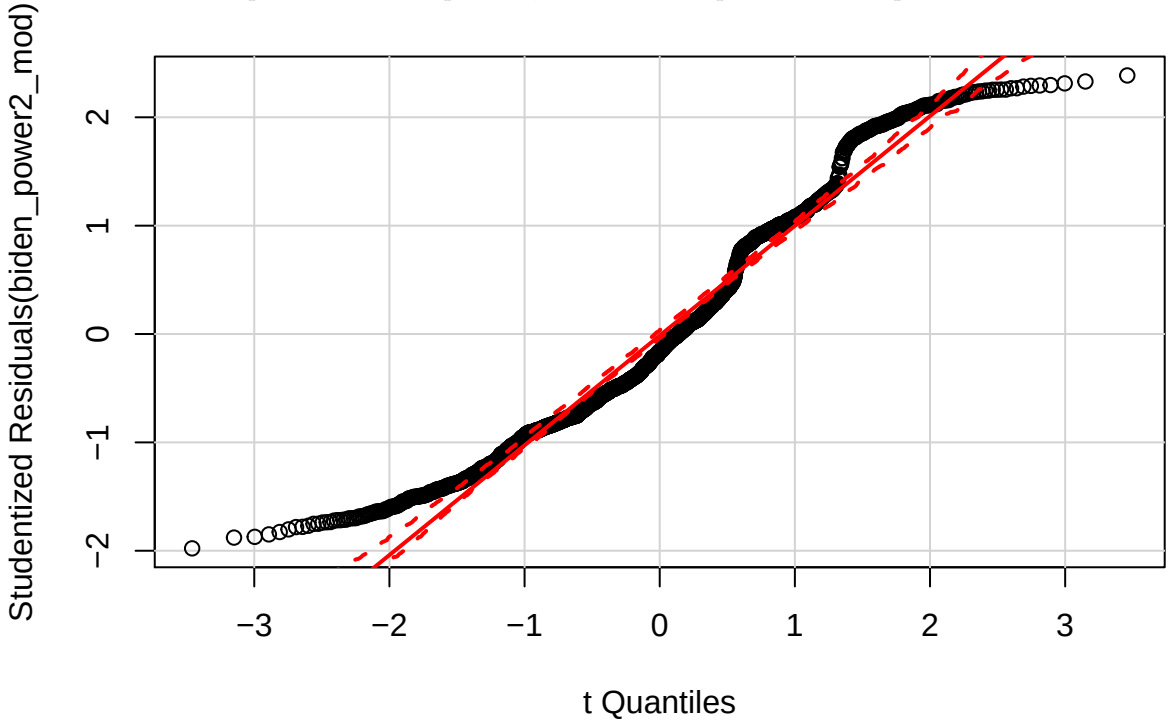


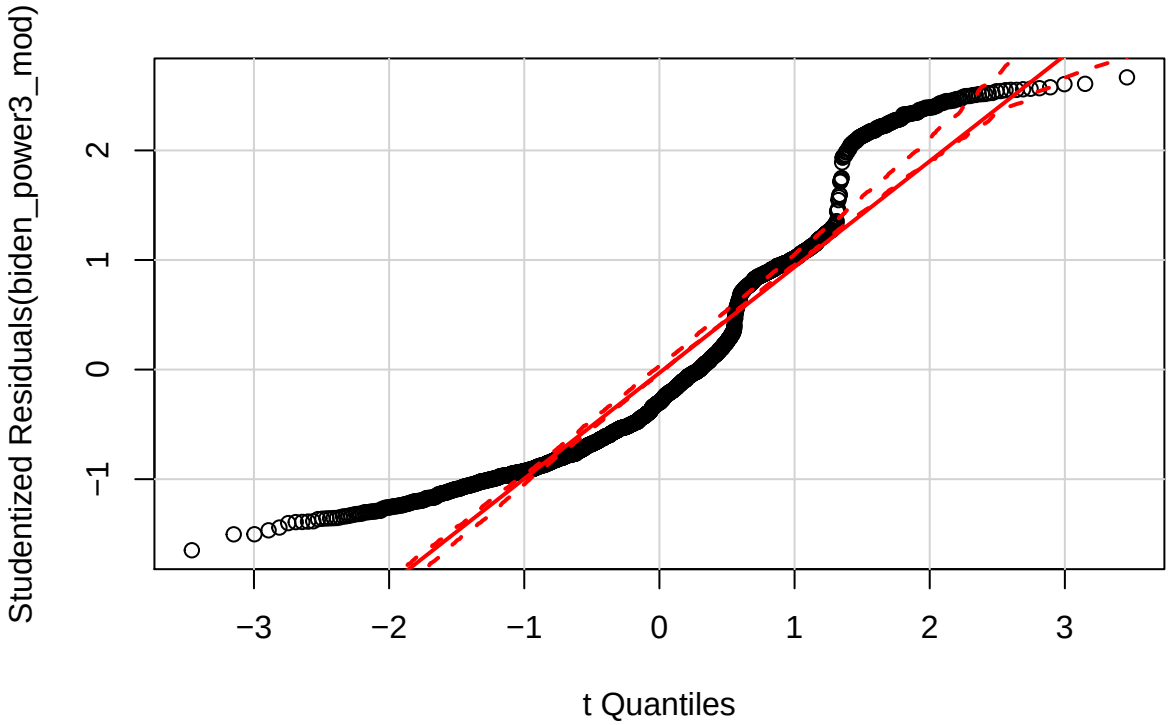
From the above histograms we can see that older age, lower biden score, male, and being a Republican seemed to be more representative in the usual/influential group. Moving forward, regarding our initial model, I think we should include party affiliation into consideration.

2. Test for non-normally distributed errors



The dashed lines indicate 95% confidence intervals calculated under the assumption that the errors are normally distributed. If any observations fall outside this range, this is an indication that the assumption has been violated. Clearly, here that is the case. Power and log transformations are typically used to correct this problem. Because not all response values are positive, so here I used power two and power three transformations.





It seems like the second power transformation is more optimal.

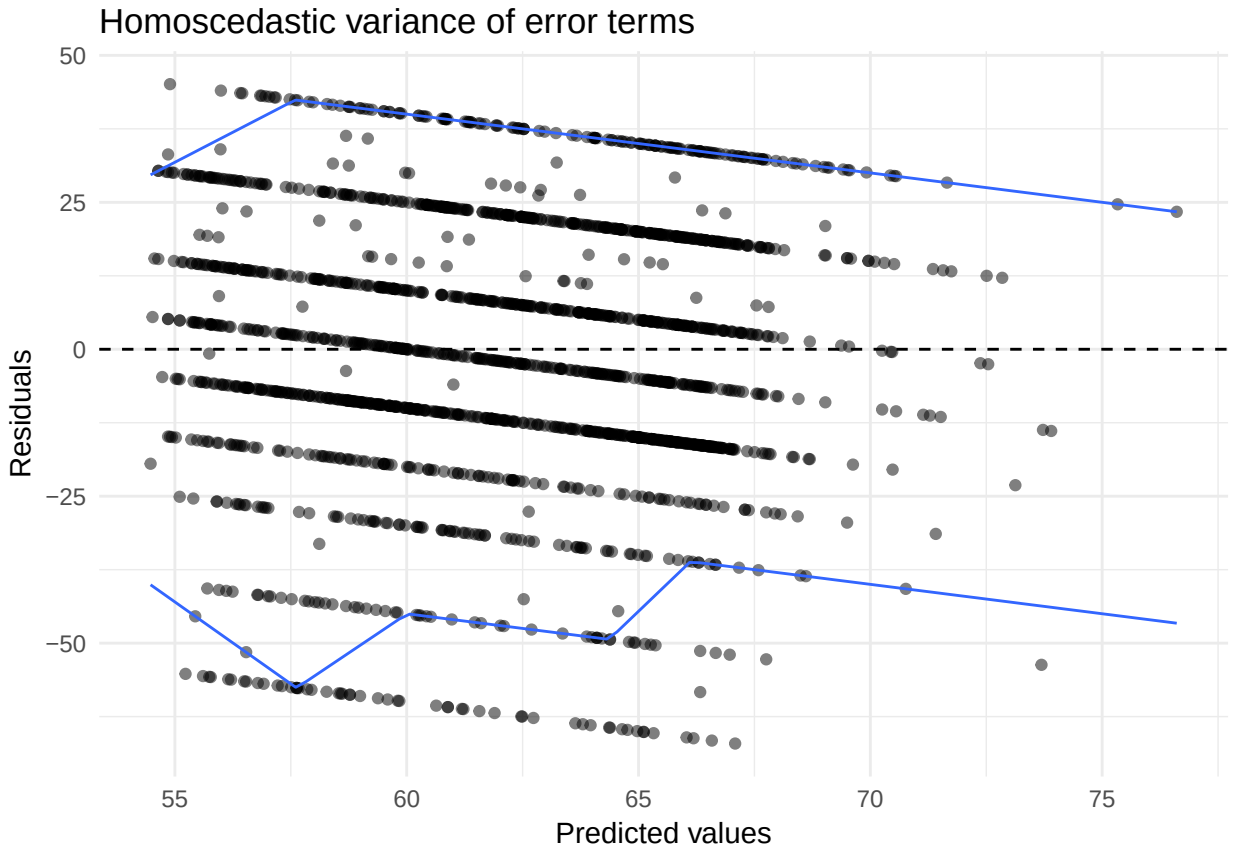
3. Test for heteroscedasticity in the model

```
## Loading required package: SparseM
```

```
##
## Attaching package: 'SparseM'
```

```
## The following object is masked from 'package:base':
##
##     backsolve
```

```
## Smoothing formula not specified. Using: y ~ qss(x, lambda = 5)
```




```
##
##  studentized Breusch-Pagan test
##
## data:  biden_mod
## BP = 20, df = 3, p-value = 5e-05
```

From the residuals vs. predicted value plot we see the spread of residuals decreases as the fitted values increase, indicating heteroscedasticity in the model. In addition, a small p-value also indicates heteroscedasticity in the model.

```
# convert residuals to weights
weights <- 1 / residuals(biden_mod)^2

biden_wls <- lm(biden ~ female + educ + age, data = biden, weights = weights)

tidy(biden_mod)
```

```
##          term estimate std.error statistic  p.value
## 1 (Intercept)  68.6210    3.5960     19.08 4.34e-74
## 2         age   0.0419    0.0325      1.29 1.98e-01
## 3        educ  -0.8887    0.2247     -3.96 7.94e-05
## 4      female   6.1961    1.0967      5.65 1.86e-08
```

```
tidy(biden_wls)
```

```
##          term estimate std.error statistic   p.value
## 1 (Intercept)  69.0173   0.33678     204.9  0.00e+00
## 2      female   5.9697   0.12875      46.4 1.20e-309
## 3        educ  -0.9098   0.02879     -31.6 9.45e-175
```

```
## 4         age   0.0388   0.00227      17.1  6.35e-61
```

We see some mild changes in the estimated parameters, but drastic reductions in the standard errors.

4. Test for multicollinearity.

```
car::vif(biden_mod)
```

```
##    age   educ female
##   1.01   1.01   1.00
```

Since no VIF statistic in the model is greater than 10, it indicates that there is no potential multicollinearity in the model.

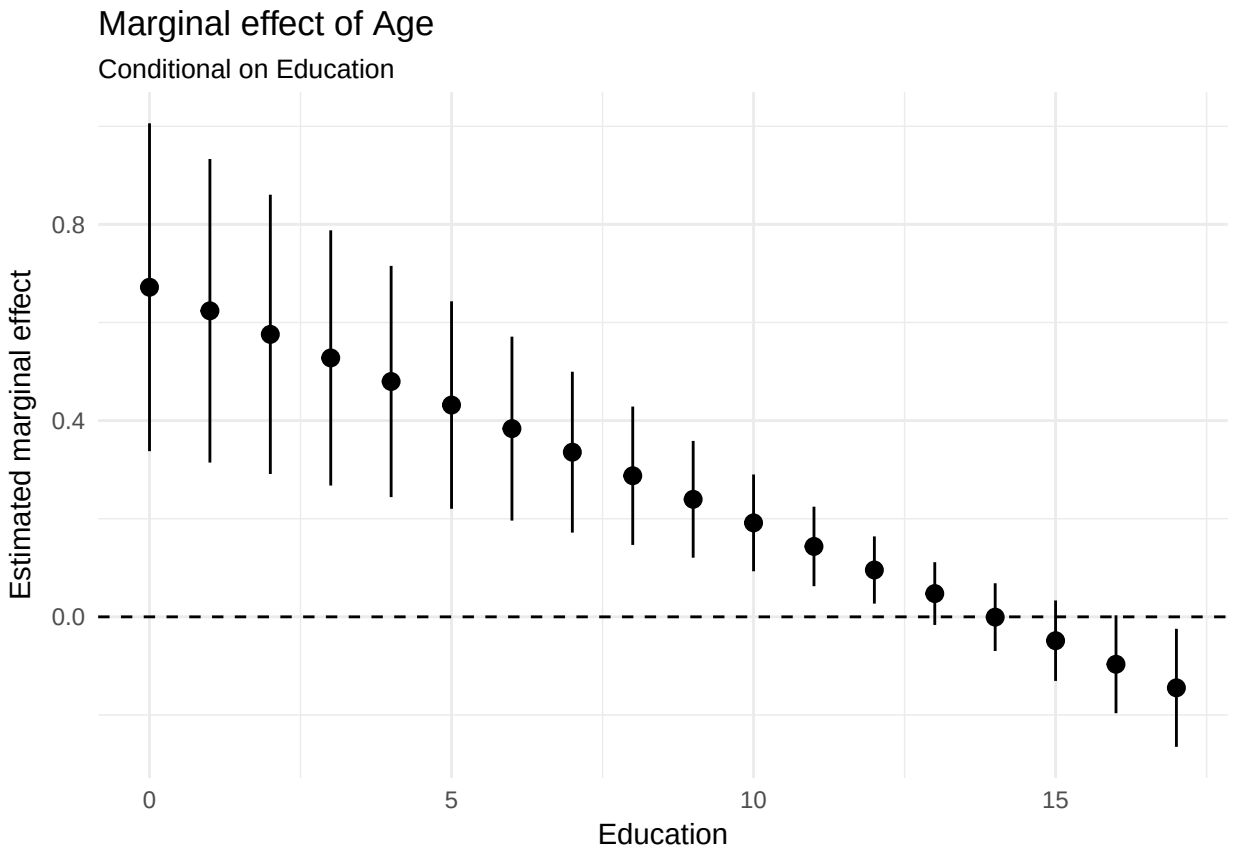
Problem 2

```
biden_second_mod <- lm(biden ~ age+educ+age*educ,data=biden)
tidy(biden_second_mod)
```

```
##          term estimate std.error statistic  p.value
## 1 (Intercept)   38.374    9.5636      4.01 6.25e-05
## 2         age    0.672    0.1705      3.94 8.43e-05
## 3        educ    1.657    0.7140      2.32 2.04e-02
## 4    age:educ   -0.048    0.0129     -3.72 2.03e-04
```

The coefficients for age, education and age:education are: 0.6719,1.6574,-0.048 and the standard errors for age, education and age:education are:0.1705,0.714,0.0129.

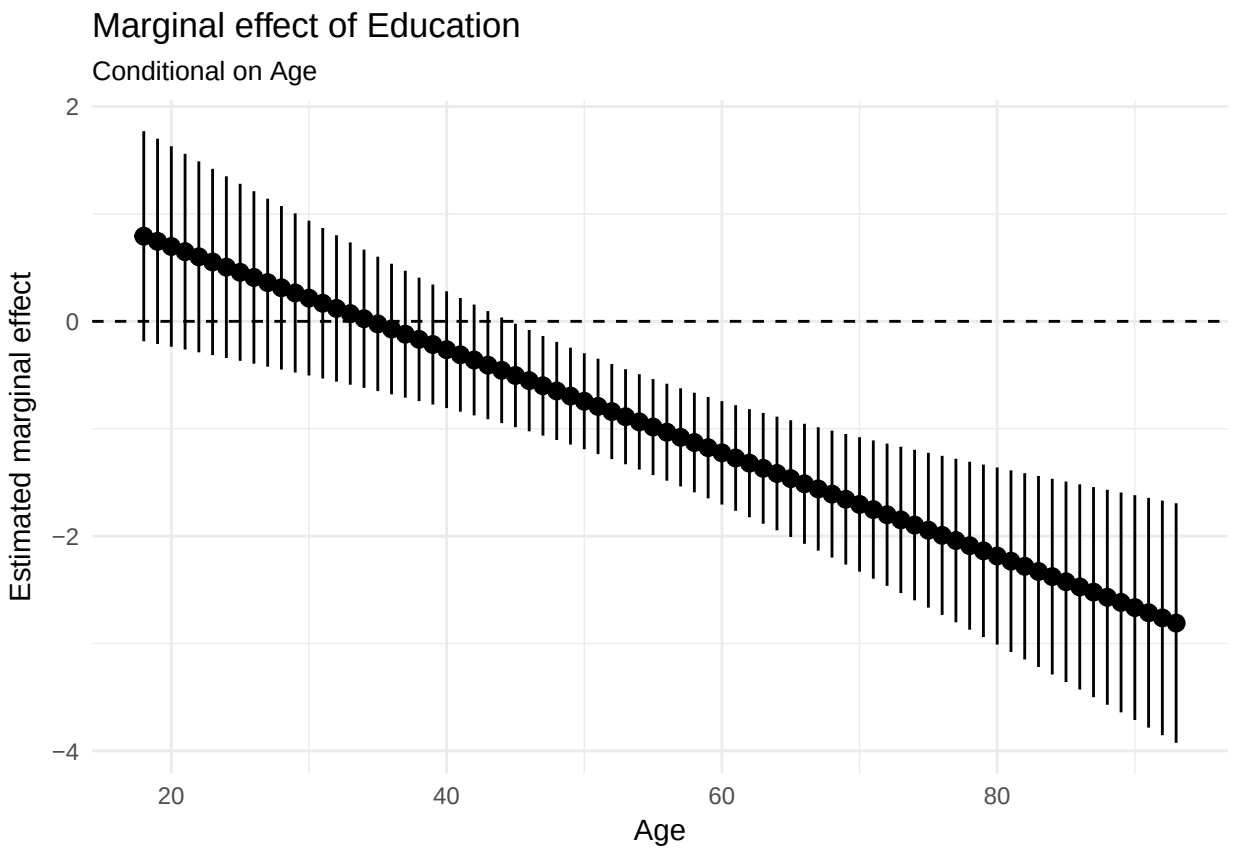
1.Marginal Effect of Age




```
## Linear hypothesis test
##
## Hypothesis:
## age  + age:educ = 0
##
## Model 1: restricted model
## Model 2: biden ~ age + educ + age * educ
##
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1   1804 985149
## 2   1803 976688  1      8461 15.6  8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the hypothesis test we get p-value below .05, therefore we can conclude that the marginal effect of age is statistically significant.The magnitude and direction can be seen from the plot.

2.Marginal Effect of Education




```
## Linear hypothesis test
##
## Hypothesis:
## educ  + age:educ = 0
##
## Model 1: restricted model
## Model 2: biden ~ age + educ + age * educ
##
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1   1804 979537
## 2   1803 976688  1      2849 5.26  0.022 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the hypothesis test we get p-value below .05, therefore we can conclude that the marginal effect of education is statistically significant.The magnitude and direction can be seen from the plot.

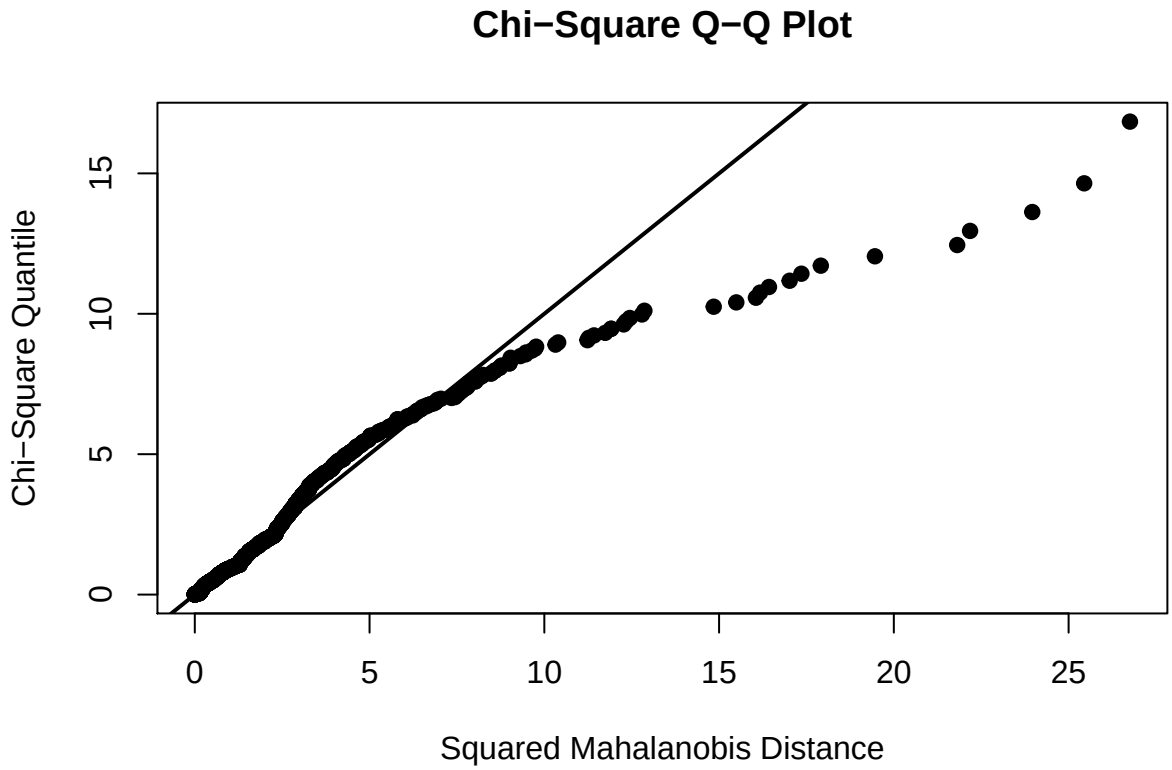
3. Missing Data

Before I use the imputation, I should test the data for multivariate normality. There are many tests in MVN packages, I chose mardiaTest.

```
library(MVN)
```

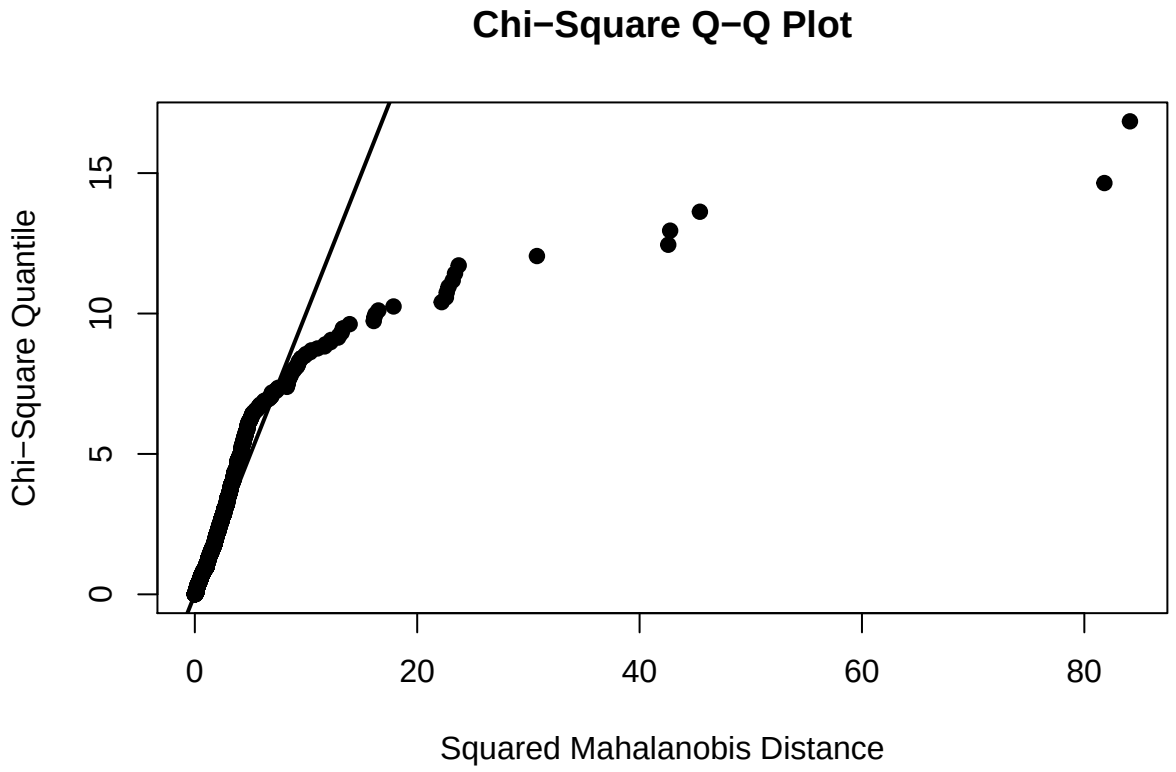
```
## sROC 0.1-2 loaded
```

```r
biden_num <- biden_data %>%
  select(educ, age)
mardiaTest(biden_num, qqplot = TRUE)
```



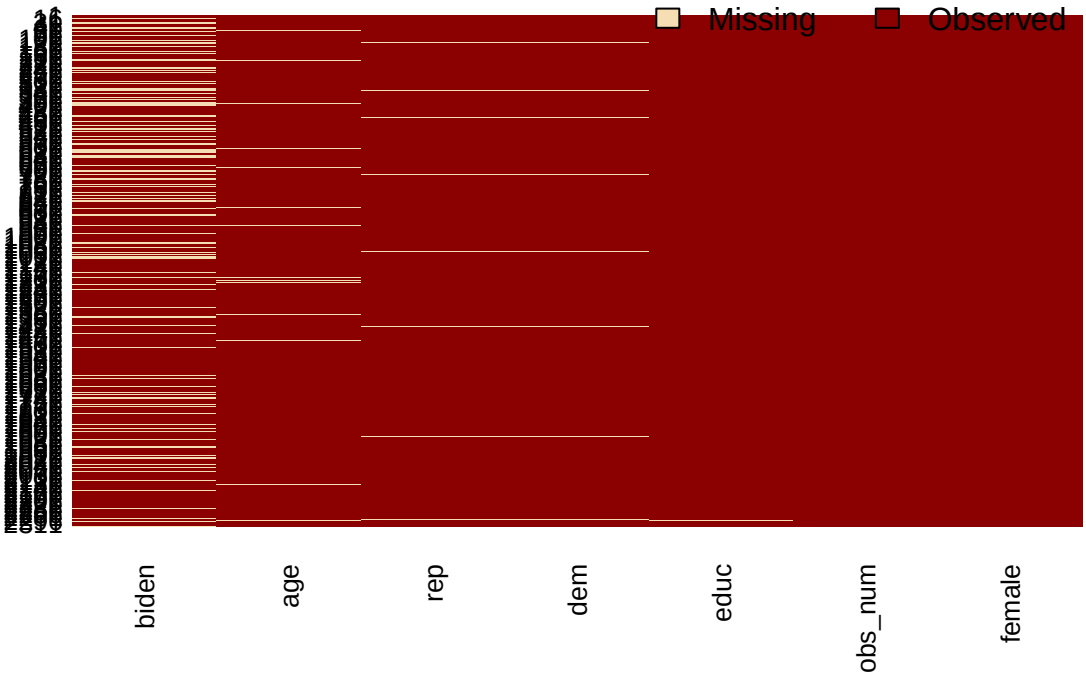

```
##    Mardia's Multivariate Normality Test
## ---------------------------------------
##    data : biden_num
##
##    g1p            : 0.9
##    chi.skew       : 341
##    p.value.skew   : 1.64e-72
##
##    g2p            : 9.21
##    z.kurtosis     : 7.21
##    p.value.kurt   : 5.61e-13
##
##    chi.small.skew : 342
##    p.value.small  : 1.13e-72
##
##    Result          : Data are not multivariate normal.
## ---------------------------------------
```

From the result from Mardia's MVN mardiaTest, we can see that the data is not multivariate normal. Let's try and use either a square root and log transformation.




```
##    Mardia's Multivariate Normality Test
## ---------------------------------------
##    data : biden_trans %>% select(sqrt_educ, sqrt_age)
##
##    g1p            : 3.91
##    chi.skew       : 1482
##    p.value.skew   : 1.26e-319
##
##    g2p            : 17.3
##    z.kurtosis     : 55.5
##    p.value.kurt   : 0
##
##    chi.small.skew : 1485
##    p.value.small  : 2.47e-320
##
##    Result          : Data are not multivariate normal.
## ---------------------------------------
```

After the squre root transformation, even though the data is still not multivariate normal, the results are better than before.




```
## # A tibble: 20 × 6
##       id        term estimate std.error statistic   p.value
##    <chr>       <chr>    <dbl>     <dbl>     <dbl>     <dbl>
## 1   imp1 (Intercept)  67.8406    2.9657    22.875 1.34e-104
## 2   imp1         age   0.0545    0.0275     1.984  4.74e-02
## 3   imp1      female   5.5425    0.9563     5.796  7.72e-09
## 4   imp1        educ  -0.8629    0.1842    -4.685  2.97e-06
## 5   imp2 (Intercept)  69.6898    2.9750    23.425 4.25e-109
## 6   imp2         age   0.0266    0.0276     0.963  3.36e-01
## 7   imp2      female   6.3394    0.9614     6.594  5.28e-11
## 8   imp2        educ  -0.9185    0.1851    -4.962  7.50e-07
## 9   imp3 (Intercept)  65.3000    2.9979    21.782  7.41e-96
## 10  imp3         age   0.0563    0.0277     2.031  4.23e-02
## 11  imp3      female   5.3107    0.9643     5.508  4.04e-08
## 12  imp3        educ  -0.6609    0.1859    -3.554  3.86e-04
## 13  imp4 (Intercept)  67.1565    3.0276    22.181  5.09e-99
## 14  imp4         age   0.0402    0.0281     1.431  1.53e-01
## 15  imp4      female   5.7220    0.9801     5.838  6.03e-09
## 16  imp4        educ  -0.7830    0.1886    -4.152  3.42e-05
## 17  imp5 (Intercept)  68.0921    2.9825    22.831 3.08e-104
## 18  imp5         age   0.0378    0.0277     1.366  1.72e-01
## 19  imp5      female   6.0896    0.9636     6.320  3.14e-10
## 20  imp5        educ  -0.8253    0.1854    -4.450  8.97e-06
```

```
## [1] "Comparison between imputed model and original model"
```

```
## Joining, by = "term"
```

```
##          term estimate std.error estimate.mi std.error.mi
## 1 (Intercept)  68.6210    3.5960     67.6158       3.4619
## 2         age   0.0419    0.0325      0.0431       0.0309
```

```
## 3        educ  -0.8887    0.2247     -0.8101       0.2142
## 4      female   6.1961    1.0967      5.8008       1.0665
```

From the above table, we can see that there does not seem to be significant difference between the coefficients of the linear model before and after the imputation process. From my opinion I think it is because the problem of non multivariate normality didn't get solved completely in previous question, so there did not show a significant change.